MILLIMETER-SCALE GENETIC GRADIENTS AND COMMUNITY-LEVEL MOLECULAR CONVERGENCE IN A HYPERSALINE MICROBIAL MAT

Victor Kunin[1], Jeroen Raes[2], J. Kirk Harris[3], John R. Spear[4], Jeffrey J. Walker[5], Natalia Ivanova[6], Christian von Mering[7], Brad M. Bebout[8], Norman R. Pace[5], Peer Bork[2] and Philip Hugenholtz[1¶].

[1] Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[2] European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

[3] Department of Pediatrics, University of Colorado Denver, Aurora, CO 80045, USA

[4] Division of Environmental Science and Engineering, Colorado School of Mines, Golden, Colorado 80401, USA;

[5] Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347, USA;

[6] Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[7] Institute of Molecular Biology, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich, Switzerland

[8] Microbial Ecology/Biogeochemistry research laboratory, NASA Ames research center, Moffett Field CA, USA

[¶]Corresponding author: fax 925-296-5720 • email: **phugenholtz@lbl.gov**

Character count (excluding Methods): 20940.

Abstract

To investigate the extent of genetic stratification in structured microbial communities, we compared the metagenomes of 10 successive layers of a phylogenetically complex hypersaline mat from Guerrero Negro, Mexico. We found pronounced millimeter-scale genetic gradients that are consistent with the physicochemical profile of the mat. Despite these gradients, all layers displayed near identical and acid-shifted isoelectric point profiles due to a molecular convergence of amino acid usage indicating that hypersalinity enforces an overriding selective pressure on the mat community.

Introduction

Ecosystems often exhibit distinct gradients. Physicochemical gradients have long been documented, but only recently has environmental shotgun sequencing allowed the associated functional (gene-based) gradients of an ecosystems biota to be addressed. Macroscale functional gradients have been inferred from oceanic metagenomic datasets, both horizontally (Johnson *et al*, 2006; Rusch *et al*, 2007; Venter *et al*, 2004) and vertically (DeLong *et al*, 2006). Many structured microbial communities have been shown to produce steep physicochemical gradients on the scale of millimeters (Jorgensen *et al*, 1979; Ley *et al*, 2006; Ludemann *et al*, 2000; Schmitt-Wagner and Brune, 1999), but associated community-level functional gradients have not been demonstrated to date.

Here, we investigate a complex, stratified, hypersaline microbial mat from Guerrero Negro, Baja California Sur, Mexico as a model for fine-scale functional variation (Ley *et al*, 2006). The dense, tofu-like texture of this mat allows intact cross-

sections to be obtained down to ~1 mm thickness. The mat shows pronounced physicochemical variation both in space and time: oxygen is detected routinely in the top 2 millimeters during the day (up to 700 μM), and the mat is completely anoxic during the night. The permanently anoxic lower layers are characterized by μM sulfide levels increasing with depth. The mat, dominated by bacteria, was reported to be one of the world's richest and most diverse microbial communities, comprising at least 752 observed species from 42 bacterial phyla, including 15 novel candidate phyla (Ley *et al*, 2006). Since the mat grows in hypersaline waters (~3X the salinity of seawater), we were also interested to look for evidence of molecular adaptations to hypersalinity in the mat community.

Results and Discussion

To investigate millimeter-scale genetic and associated functional stratification, we performed a metagenomic analysis of 10 spatially successive layers of the Guerrero Negro mat. Mat core samples were collected during the day (Table S1) and upper layers were sectioned at a finer scale (1 mm slices) than the lower layers (4 to 15 mm slices) to capture variation associated with the steep oxygen gradient in the upper millimeters of the mat (Table S2). DNA from each layer was cloned and shotgun-sequenced using capillary sequencing with an average of ~13,000 reads per layer. No significant assembly of the reads was possible, even when all data were combined (largest contig was 8.4 kb from a combined assembly). We chose therefore to analyze only the unassembled data (average trimmed (Chou and Holmes, 2001) read length 808 bp) to avoid chimerism that has been reported to be frequent in contigs <10 kb (Mavromatis *et al*, 2007). Genes were

predicted on vector and quality-trimmed reads with fgenesb (http://www.softberry.com/) using a generic bacterial model, resulting in an average of 13,600 genes per layer (Table S2). These data are available through the IMG/M system (Markowitz *et al*, 2006) at http://durian.jgi-psf.org/cgi-bin/img_mi_v240/main.cgi (username/password: public/public).

Using both bulk similarity matches and phylogenetic mapping of conserved marker genes (von Mering *et al*, 2007a), we found strong phylogenetic variation between layers. Cyanobacteria and Alphaproteobacteria were the most abundant lineages in the top two layers (Fig. S1). Below the upper 2 mm, Proteobacteria, Bacteroidetes, Chloroflexi and Planctomycetes were the most represented phyla, with a notable peak in Bacteroidetes at 3 mm (Fig. S1). Numerous traces of other bacterial phyla as well as some archaea and eukaryotes were also identified. A large fraction of predicted proteins in layers below 2 mm did not have significant sequence similarity to any protein in public databases, reflecting the high degree of phylum-level novelty in the mat community (Ley *et al*, 2006). These metagenome-based findings are in broad agreement with single marker gene surveys of the mat (Ley *et al*, 2006; Spear *et al*, 2003).

A rough measure of functional potential per organism can be made by estimating the average effective genome size (EGS) (Raes *et al*, 2007). Using this method, we predicted an increased average bacterial genome size at the border of the oxic and anoxic zone (1-2 mm depth); 6 Mb at the border vs 3-3.5 Mb for the rest of the mat (Fig. S2). This may reflect an increased functional complexity needed for survival in the constantly fluctuating conditions at this depth as was recently observed in the genome of a marine *Beggiotoa* occupying a similar niche (Mussmann *et al*, 2007) .

To investigate genetic gradients through the mat, we determined the relative abundances of individual gene families and metabolic pathways between mat layers, and compared the mat data to external datasets for reference. Many gene families were highly abundant in the mat despite high overall functional diversity (Fig. S3) and very low sequence coverage of individual species. Indeed, the mat dataset roughly doubled existing inventories for some of the gene families described below (Table 1). This implies that multiple species and likely higher-level taxa contribute representatives of these families, and suggests that there has been strong selection for a limited number of common functionalities in the mat.

The key aspect of this study was to use the metagenomic data to determine what, if any, millimeter-scale genetic gradients are detectable in this very complex and structured ecosystem. Several gene families and pathways either directly (Fig. 1A) or inversely (Fig. 1B) tracked the steep oxygen gradient in the top 2 mm of the mat and sulfide gradient below 2 mm. Genes directly involved in photosynthesis (KEGG map 00195) were statistically overrepresented in the top two layers relative to lower layers. In addition, an uncharacterized protein domain (pfam05685) highly paralogous in phototrophic lineages (most cyanobacterial and some Chloroflexi genomes) showed a steep declining gradient in the top 6 mm (Fig. 1A) consistent with dominance of phototrophs in the same region. Chaperones similarly tracked the oxygen gradient when all gene families with chaperone activity are combined together. The over-representation of chaperones in the top 2 mm relative to the rest of the mat may not be associated with oxygen concentration, but rather with heat stress caused by direct exposure to sunlight.

Gene families and pathways that tracked inversely with oxygen concentration included ferredoxins, trimethylamine methyltransferase (*Mttb*), sulfatases and sugar degradation pathways (Fig. 1B). Ferredoxins and associated proteins show a four fold increase from the top layer down to a depth of 4 mm and thereafter are uniformly over-represented. Two COG families are chiefly responsible for this trend: COG1148 (heterodisulfide reductase, subunit A and related polyferredoxins) and COG2414 (Aldehyde:ferredoxin oxidoreductase). The expansion of ferredoxins in the anoxic layers likely reflects the diversification of redox reactions required for anaerobic respiration. *Mttb* (pfam06253, COG5598) methyltransferase does not become significantly over-represented until at least 7 mm into the mat (Fig. 1B), well below the anoxic boundary. *Mttb* was initially identified as a protein facilitating the first step of methanogenesis from trimethylamine in *Methanosarcinaceae* (Paul *et al*, 2000). However, this gene family is also found in methylotrophic bacteria (e.g. in *Rhodobacteraceae* and *Rhizobiaceae*), suggesting a more generalized role in C1 metabolism.

One of the most pronounced inverse gradients is observed for sulfatases (COG3119) that are involved in hydrolysis of sulfated organic compounds (Fig. 1B). Since sulfatases can function in the presence of oxygen, the gradient is presumably a reflection of availability of sulfated compounds in the mat. While the concentration gradient of sulfated compounds is not known in the mat, they are produced by phototrophs (Kates, 1986) and are widespread in marine environments (Glockner *et al*, 2003). Sulfatase genes obtained from the mat exhibited extensive sequence divergence suggesting that a corresponding wide variety of sulfated organic substrates are present in the mat with the highest concentrations below 2 mm. The over-representation of this gene

family may in part be due to an expansion of sulfatase genes in the genomes of planctomycetes, suggested to be involved particularly in hydrolysis of sulfated glycopolymers (Glockner *et al*, 2003).

Sugar degradation pathways (glycolysis, pentose and uronic acid degradation) show a two-fold increase with depth through the top 3 mm and maintain high relative representation in the anoxic lower layers (Fig. 1B). This suggests that heterotrophic metabolism of sugars, particularly pentoses and uronic acids, is important in the lower layers.

Organisms living at the boundary between the oxic and anoxic zones could potentially accumulate substrates with high reductive potential in the anoxic zone, and then move to the oxic zone to harvest this potential by oxidation (Mussmann *et al*, 2007). This would require boundary zone organisms to be motile and chemotactic. Indeed, we find that chemotaxis signature genes peak sharply at the oxic-anoxic boundary (Fig. 1C). Flagella appear not to be the dominant source of motility in these chemotactic organisms as flagellar genes actually dip in this region (Fig. 1C). Chemotactic gliding bacteria have been observed in fresh mat cores (Garcia-Pichel *et al*, 1994; Kruschel and Castenholz, 1998) and our molecular data suggest they are most abundant in the boundary zone, bridging the oxic and anoxic layers.

Despite the pronounced phylogenetic and functional gradients in the mat, hypersalinity is a selective pressure common to the whole community. A known adaptation to hypersalinity is enrichment of proteins with acidic amino acids allowing proteins to function in high cytoplasmic salt concentrations (Soppa, 2006). The resulting acid-shifted protein isoelectric points have been documented in the genomes of only two

lineages, the archaeal class Halobacteria (Kennedy *et al*, 2001; Soppa, 2006) and the bacterial species *Salinibacter ruber* (Mongodin *et al*, 2005; Oren and Mana, 2002), so it is unclear how widespread this mechanism is in halophilic communities.

The average isoelectric points of the mat layer communities are conspicuously acid-shifted compared to most bacteria and microbiomes that are non-halophilic (Fig. 2A). We determined this to be due primarily to an enrichment in the acidic amino acid, aspartate (Fig. 2B). Furthermore, the isoelectric profiles of all 10 layers converge on a common acid-shifted profile (Fig. 3A) despite a significant variation in GC content between layers (Fig. 3B), reflecting differing phylogenetic composition. The latter is consistent with aspartate usage being GC-independent since it can be encoded by both GC-rich and -poor codons (GAC and GAT respectively). As each metagenomic read pair likely is derived from different species and no single species dominates the mat community, we conclude that a significant fraction of the community has converged on the enrichment of low isoelectric point proteins.

In summary, this study demonstrates that millimeter-scale genetic gradients can be readily discerned through a vertical cross-section of a highly structured and complex microbial community using low sequence coverage. Further, we could directly and inversely correlate many of the genetic gradients to the physicochemical profile of the mat. Microbial biofilms are important in many habitats, including our own bodies (Eckburg *et al*, 2005; Kroes *et al*, 1999) and often display physicochemical gradients at mm to cm scales. However few biofilms are as robust as microbial mats and methods may need to be adapted to preserve spatial structure (Webster *et al*, 2006) and allow the relevant fine-scale genetic gradients to be resolved.

Surprisingly, we found that adaptation to hypersalinity by enriching proteins with acidic amino acids is more widespread than previously appreciated. While this is the first example of species-independent molecular convergence in a microbial community, we predict that similar convergence patterns will be observed in other communities adapted to similar or different environmental conditions, such as temperature (Gianese *et al*, 2001) or pressure (Lauro and Bartlett, 2008; Simonato *et al*, 2006).

Methods

Mat core samples were collected around 2 pm from pond 4 near 5 at the Exportadora de Sal Saltworks, Guerrero Negro, Baja California Sur, Mexico. The salinity of the bulk water above the mat was ~9% (~3X the salinity of seawater). Other metadata for the sample can be found in Table S1. Four replicate cores were collected, sectioned into layers with sterile scalpals and DNA extracted, normalized, pooled and sequenced as described in Supporting Information.

Community composition analysis was performed using the consensus of i) best BlastP hits (Altschul *et al*, 1997) to the IMG/M database (Markowitz *et al*, 2006) and ii) phylogenetic mapping of signature genes on a phylogenetic tree (von Mering *et al*, 2007a). See Supporting Information for details.

Gene-based functional gradients were calculated as follows: genes were assigned to their COG families (Tatusov *et al*, 1997) and pfam domains (Bateman *et al*, 2002) based on rpsBLAST (Altschul *et al*, 1997). The gradients were examined for possible over-representation of groups or individual families or domains, and 1000 bootstrap iterations were used to assess the significance of over-representation. The described gradients were independently confirmed using two databases; IMG/M (Markowitz *et al*,

2006) and the STRING database (von Mering *et al*, 2007b). Further details, as well as groupings of families/domains are described in Supporting Information.

Isoelectric point distributions, amino acid composition, and GC content were computed using appropriate perl scripts and modules as described in Supporting Information.

Acknowledgements

We thank Amber Hartman for fruitful discussions and the Exportadora de Sal Saltworks in Guerrero Negro, Baja California, Sur for access and assistance with the field site. We also thank the NASA Funded researchers at NASA Ames who assist with permitting and field work; David DesMarias, Moira Doty, Tori Hoehler, Mary Hogan, and Kendra Turk. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. JR and PB are supported by the European Union 6th Framework Program (Contract No. LSHG-CT-2004-503567).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25:** 3389-3402.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. *Nucleic acids research* **30:** 276-280.

Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics (Oxford, England)* **17:** 1093-1104.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science (New York, NY* **311:** 496-503.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. *Science (New York, NY* **308:** 1635-1638.

Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24:** 1263-1269.

Garcia-Pichel F, Mechling M, Castenholz RW (1994) Diel migrations of microorganisms within a benthic, hypersaline mat community. *Appl Environ Microbiol* **60:** 1500-1511.

Gianese G, Argos P, Pascarella S (2001) Structural adaptation of enzymes to low temperatures. *Protein engineering* **14:** 141-148.

Glockner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R (2003) Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 8298-8303.

http://www.softberry.com/ SoftBerry - fgenesb.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science (New York, NY* **311:** 1737-1740.

Jorgensen BB, Revsbech NP, Blackburn TH, Cohen Y (1979) Diurnal cycle of oxygen and sulfide microgradients and microbial photosynthesis in a cyanobacterial mat sediment. *Appl Environ Microbiol* **38:** 46-58.

Kates M (1986) *Techniques of lipidology : isolation, analysis, and identification of lipids*, 2nd rev. edn. Amsterdam, New York: Elsevier.

Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S (2001) Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res* **11:** 1641-1650.

Kroes I, Lepp PW, Relman DA (1999) Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences of the United States of America* **96:** 14547-14552.

Kruschel C, Castenholz R (1998) The effect of solar UV and visible irradiance on the vertical movements of cyanobacteria in microbial mats of hypersaline waters. *FEMS Microbiology Ecology* **27:** 53-72.

Lauro FM, Bartlett DH (2008) Prokaryotic lifestyles in deep sea habitats. *Extremophiles* **12:** 15-25.

Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, Maresca JA, Bryant DA, Sogin ML, Pace NR (2006) Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol* **72:** 3685-3695.

Ludemann H, Arth I, Liesack W (2000) Spatial changes in the bacterial community structure along a vertical oxygen gradient in flooded paddy soil cores. *Appl Environ Microbiol* **66:** 754-762.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavrommatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P, Kyrpides NC (2006) An experimental metagenome data management and analysis system. *Bioinformatics (Oxford, England)* **22:** e359-367.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides NC (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4:** 495-500.

Mongodin EF, Nelson KE, Daugherty S, Deboy RT, Wister J, Khouri H, Weidman J, Walsh DA, Papke RT, Sanchez Perez G, Sharma AK, Nesbo CL, MacLeod D, Bapteste E, Doolittle WF, Charlebois RL, Legault B, Rodriguez-Valera F (2005) The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 18147-18152.

Mussmann M, Hu FZ, Richter M, de Beer D, Preisler A, Jorgensen BB, Huntemann M, Glockner FO, Amann R, Koopman WJ, Lasken RS, Janto B, Hogg J, Stoodley P, Boissy R, Ehrlich GD (2007) Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS biology* **5:** e230.

Oren A, Mana L (2002) Amino acid composition of bulk protein and salt relationships of selected enzymes of Salinibacter ruber, an extremely halophilic bacterium. *Extremophiles* **6:** 217-223.

Paul L, Ferguson DJ, Jr., Krzycki JA (2000) The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of Methanosarcina barkeri contain in-frame and read-through amber codons. *J Bacteriol* **182:** 2520-2529.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* **8:** R10.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The Sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS biology* **5:** e77.

Schmitt-Wagner D, Brune A (1999) Hydrogen profiles and localization of methanogenic activities in the highly compartmentalized hindgut of soil-feeding higher termites (Cubitermes spp.). *Appl Environ Microbiol* **65:** 4490-4496.

Simonato F, Campanaro S, Lauro FM, Vezzi A, D'Angelo M, Vitulo N, Valle G, Bartlett DH (2006) Piezophilic adaptation: a genomic point of view. *Journal of biotechnology* **126:** 11-25.

Soppa J (2006) From genomes to function: haloarchaea as model organisms. *Microbiology* **152:** 585-590.

Spear JR, Ley RE, Berger AB, Pace NR (2003) Complexity in natural microbial ecosystems: the Guerrero Negro experience. *The Biological bulletin* **204:** 168-173.

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science (New York, NY* **278:** 631-637.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. *Science (New York, NY* **308:** 554-557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, NY* **304:** 66-74.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P (2007a) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science (New York, NY* **315:** 1126-1130.

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007b) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic acids research* **35:** D358-362.

Webster P, Wu S, Gomez G, Apicella M, Plaut AG, St Geme JW, 3rd (2006) Distribution of bacterial proteins in biofilms formed by non-typeable Haemophilus influenzae. *J Histochem Cytochem* **54:** 829-842.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, Szeto E, Kyrpides NC, Mussmann M, Amann R, Bergin C, Ruehland C, Rubin EM, Dubilier N (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443:** 950-955.

Figure legends

Figure 1. Gradients of gene families or groups of functionally-related gene families enriched in the oxic zone (A), anoxic (high H_2S) zone (B) and varying across the oxic-anoxic border (low H_2S) zone (C). Relative abundance is normalized by average number of genes in a layer. In most cases, these genes and groups of genes were over-represented relative to other metagenomic datasets (Table 1). Error bars denote standard deviations calculated from 1000 bootstrap resamplings of predicted proteins, and points with non-overlapping error bars are treated as significantly different. Lists of gene families used in each group (Photosynthesis-related proteins, Chaperones, Ferredoxins and associated proteins, Sugar degradation pathways, Chemotaxis and Flagella), as well as details of the resampling procedure are given in Supporting Information.

Figure 2. Average isoelectric point (A) and aspartate content (B) of all predicted proteins in the mat layer communities and reference bacteria, archaea, phages and microbiomes available through IMG/M (Markowitz *et al*, 2006). Genomic average was computed for each genome or microbiome, with 10 layers of the mat treated separately. These values were rounded up to the next (larger value) bin in increments of 0.2 and 0.5 in (A) and (B) respectively, and the distribution of the bins plotted as a fraction of each dataset.

Figure 3. Isoelectric point profiles of predicted proteins (A) and GC content profiles of reads (B) for mat layer communities. In A, isoelectric point profiles for selected reference genomes are added to highlight the highly overlapping and acid-shifted mat layer profiles.

Table 1. Most prominent gene families and domains in the Guerrero Negro hypersaline mat core relative to other sequenced microbiome samples[a]. Numbers represent raw counts and numbers in parentheses are normalized for mat dataset size.

Gene family or domain[b]	Annotation	Mat	AMD	Soil	Whalefall	Gutless Worm	Sludge	IMG
COG3119	Arylsulfatase A and related enzymes	640 (640)	0	195 (145)	46 (165)	16 (77)	32 (127)	1154 (55)
COG5598	Trimethylamine:corrinoid methyltransferase	112 (112)	0	16 (12)	5 (18)	52 (249)	3 (12)	114 (5)
COG1148	Heterodisulfide reductase, subunit A and related polyferredoxins	172 (172)	0	16 (12)	5 (18)	40 (192)	0	185 (9)
COG2414	Aldehyde:ferredoxin oxidoreductase	110 (110)	0	20 (15)	4 (14)	39 (187)	5 (20)	225 (11)
Pfam05685	DUF820 domain	142(142)	3 (32)	63 (47)	0	8 (38)	10 (40)	825 (40)

a. Mat (combined data from all layers, present study), AMD (acid mine drainage biofilm (Tyson *et al*, 2004)), soil (Tringe *et al*, 2005), whalefall (sample 3 (Tringe *et al*, 2005)), gutless worm (Woyke *et al*, 2006), sludge (US, (Garcia Martin *et al*, 2006)), IMG (version 2.20, combined data from 728 microbial genomes (Markowitz *et al*, 2006)).

b. COG - cluster of orthologous genes (Tatusov *et al*, 1997), pfam (Bateman *et al*, 2002)

Figure 1.

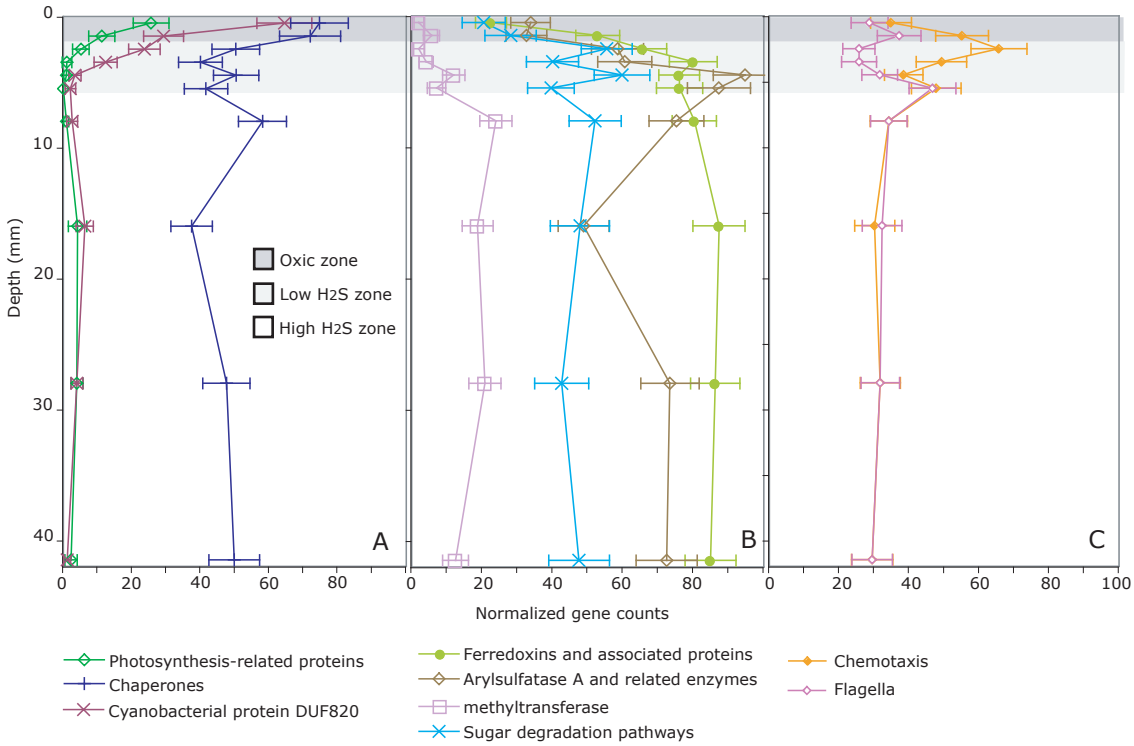

Depth (mm)
0
10
20
30
40
Oxic zone
Low H2S zone
High H2S zone
A
B
C
0
20
40
60
80
100
Normalized gene counts
Photosynthesis-related proteins
Chaperones
Cyanobacterial protein DUF820
Ferredoxins and associated proteins
Arylsulfatase A and related enzymes
methyltransferase
Sugar degradation pathways
Chemotaxis
Flagella

Figure 2.

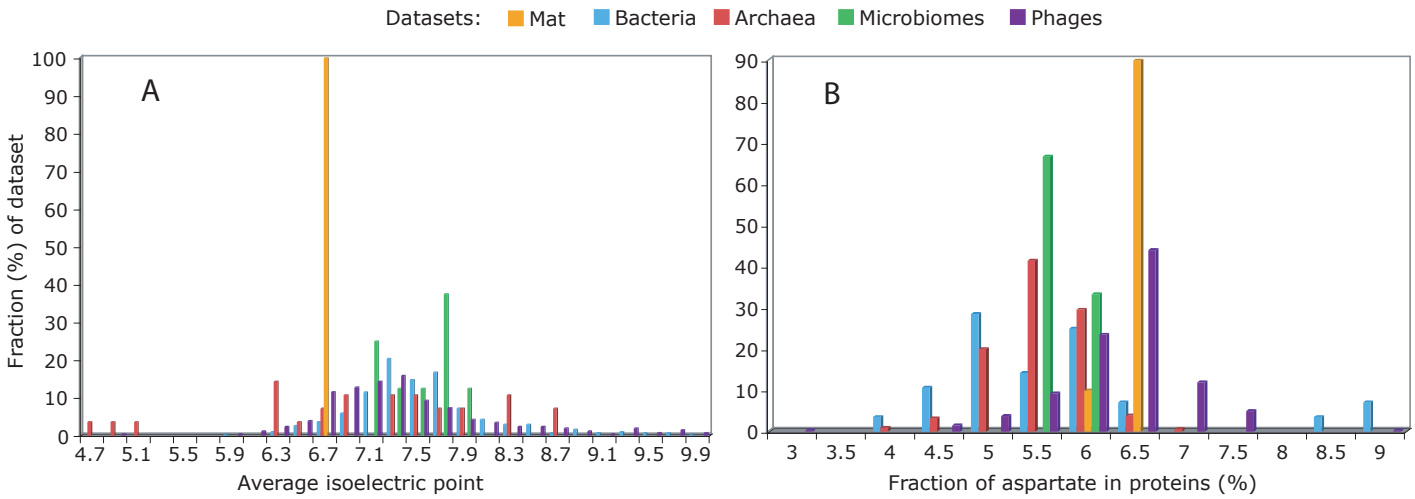
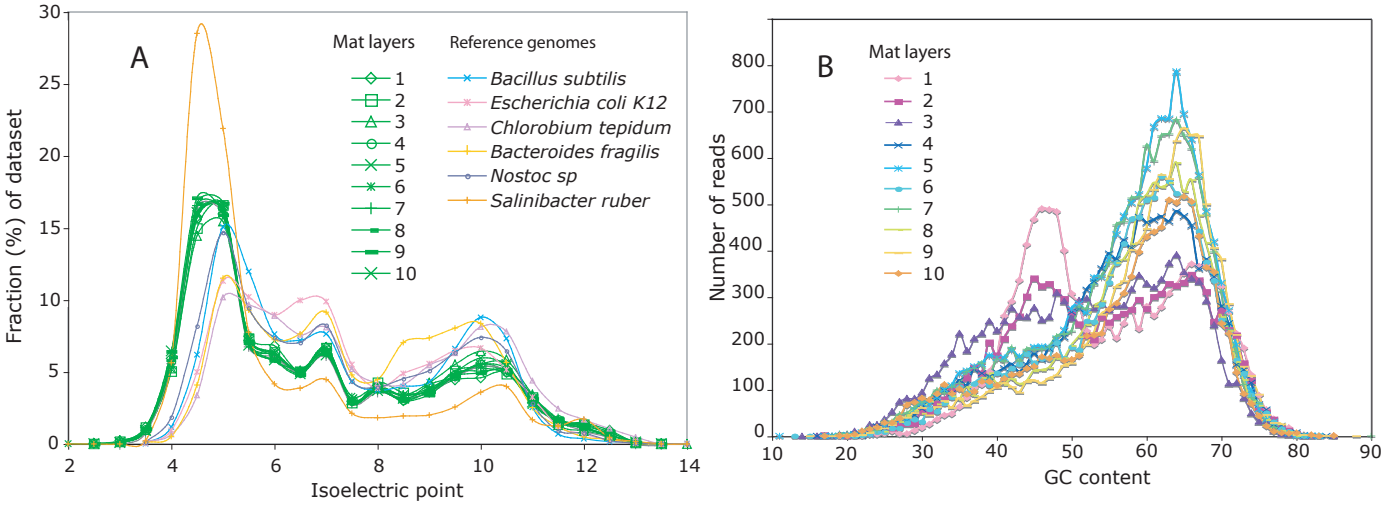

Datasets: Mat Bacteria Archaea Microbiomes Phages
A
Fraction (%) of dataset
Average isoelectric point
B
Fraction of aspartate in proteins (%)

Figure 3.


A
Mat layers
1
2
3
4
5
6
7
8
9
10
Reference genomes
Bacillus subtilis
Escherichia coli K12
Chlorobium tepidum
Bacteroides fragilis
Nostoc sp
Salinibacter ruber
Fraction (%) of dataset
Isoelectric point
B
Mat layers
1
2
3
4
5
6
7
8
9
10
Number of reads
GC content

MILLIMETER-SCALE GENETIC GRADIENTS AND COMMUNITY-LEVEL MOLECULAR CONVERGENCE IN A HYPERSALINE MICROBIAL MAT

Supporting Information

Victor Kunin[1], Jeroen Raes[2], J. Kirk Harris[3], John R. Spear[4], Jeffrey J. Walker[5], Natalia Ivanova[6], Christian von Mering[7], Brad M. Bebout[8], Norman R. Pace[5], Peer Bork[2] and Philip Hugenholtz[1¶].

[1] Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[2] European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

[3] Department of Pediatrics, University of Colorado Denver, Aurora, CO 80045, USA

[4] Division of Environmental Science and Engineering, Colorado School of Mines, Golden, Colorado 80401, USA;

[5] Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347, USA;

[6] Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA.

[7] Institute of Molecular Biology, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich, Switzerland

[8] Microbial Ecology/Biogeochemistry research laboratory, NASA Ames research center, Moffett Field CA, USA

[¶]Corresponding author: fax 925-296-5720 • email: phugenholtz@lbl.gov

Sample preparation and sequencing

A sample of hypersaline mat was collected ~100 m off-shore around 2 pm from pond 4 near 5 at the Exportadora de Sal Saltworks, Guerrero Negro, Baja California Sur Mexico (Ley *et al*, 2006) by JS (Table S1). The 25 x 25 cm x 6 cm mat piece was brought to shore in a pan of its own water. Four replicate 6 cm thick x 8 mm diameter cores were excised from the middle of the mat sample using a sterile coring instrument. Each core was sectioned into coins as described in Table S2 using sterile scalpal blades for each layer. The mat is vertically striated and these striations were used to ensure that coins were obtained from the same layer of each replicate core. Coins were frozen on liquid nitrogen. DNA was extracted by bead-beating (Ley *et al*, 2006) from individual sub-samples (~10 mg) of each mat slice. DNA for each of the 23 sections was extracted from duplicate sub-samples of each corresponding layer of each core, and combined. The average DNA yield per layer was ~3.6 μg DNA / mg of mat.

Table S1. Metadata for Guerrero Negro hypersaline mat sample used in this study.

Parameter	Value
Collection date	13 Feb 2005
Collection time	~2 pm
GPS coodinates	N27 41.345 W113 55.027
Ambient temperature	15˚C
pH	6 to 9 (varies diurnally and through layers)
Sampling depth	~ 1 m below water level
Salinity of bulk water	90 ppt
Sulfate concentration of bulk water	80 mM

Table S2. Data collection, sequencing and gene calling summary.

Layer #	Depth (mm)	Average depth (mm)	Pooled coins	Reads #	Bases #	Genes #
1	0-1	0.5		12218	8596350	13422
2	1-2	1.5		11574	7469572	12210
3	2-3	2.5		12419	8286576	13385
4	3-4	3.5		12824	8215056	13388
5	4-5	4.5		15663	9803980	16093
6	5-6	5.5		12531	8377132	13534
7	6-10	8	4x1mm	15060	9864533	16079
8	10-22	16	4x3mm	12693	8017278	13217
9	22-34	28	4x3mm	12528	8382678	13855
10	34-49	41.5	5x3mm	11637	7240715	12135

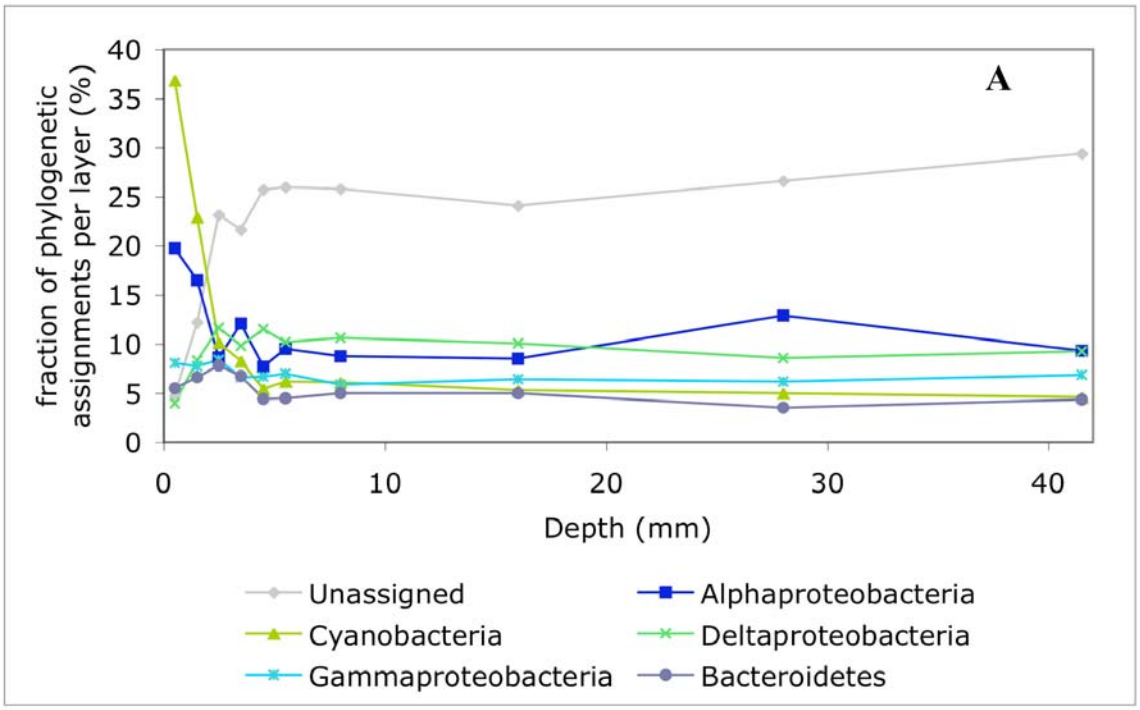
DNA for all selected layers was blunt-end ligated into pUC18 (3 kb) and end-sequenced using capillary sequencing (see Table S2 for number of reads and bases per layer). Genes were predicted on vector and quality-trimmed reads with fgenesb using a generic bacterial model (http://www.softberry.com/), resulting in an average of 13,600 genes per layer (Table S2). The data was loaded into the Integrated Microbial Genomes with Microbiome samples (IMG/M) system (Markowitz *et al*, 2006) at http://durian.jgi-psf.org/cgi-bin/img_mi_v240/main.cgi (username/password: public/public). In addition, the data was loaded into the STRING database (von Mering *et al*, 2007b).

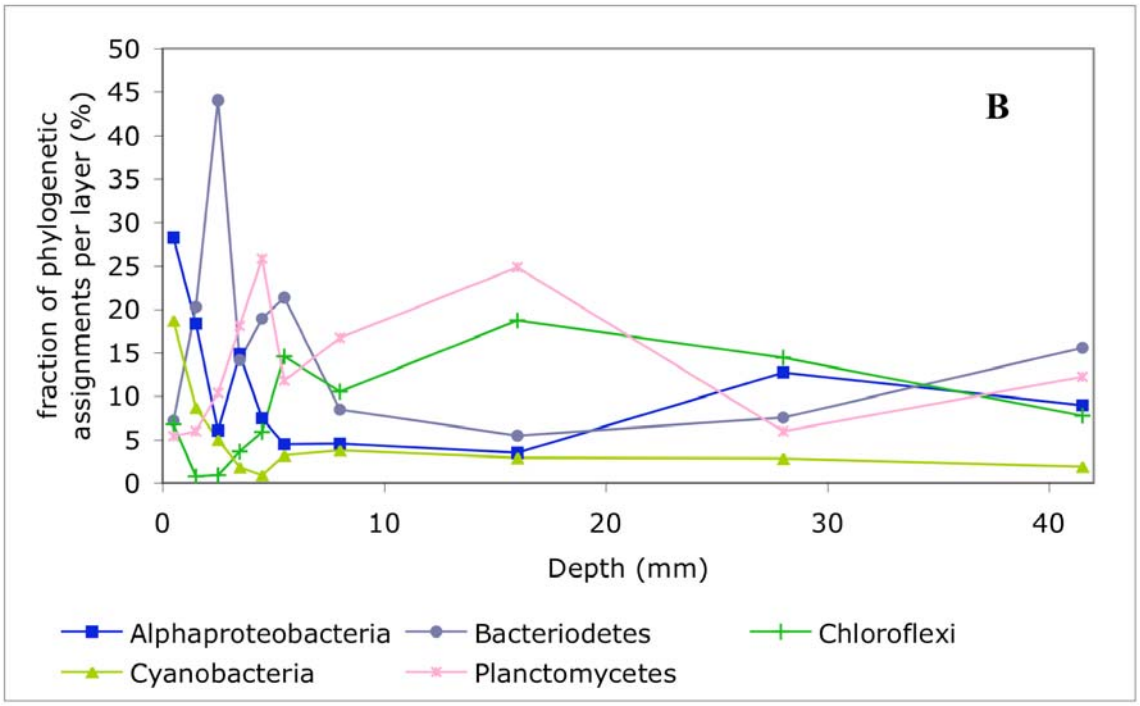
Community composition analysis

We analyzed the phylogenetic content of the mat to confirm the previously observed community composition based on a 16S rRNA gene survey (Ley *et al*, 2006). The number of 16S rRNA gene sequences in the metagenomic dataset varied from 1 to 12 per layer and therefore did not provide sufficient information for statistically significant exploration of phylogenetic distribution in the mat. We therefore analyzed the phylogenetic distribution of predicted genes based on best BLAST hits of all genes (Markowitz *et al*, 2006) (FigS1a) and phylogenetic mapping of 31 marker genes against a reference concatenated gene tree which has been shown to be quantitative in complex samples of similar size (von Mering *et al*, 2007a) (Fig S1b). For the BLAST analysis, we assessed the distribution of best BLAST hits over 30% identity to phylogenetic groups in the IMG database. IMG was chosen because all proteins in this database belong to

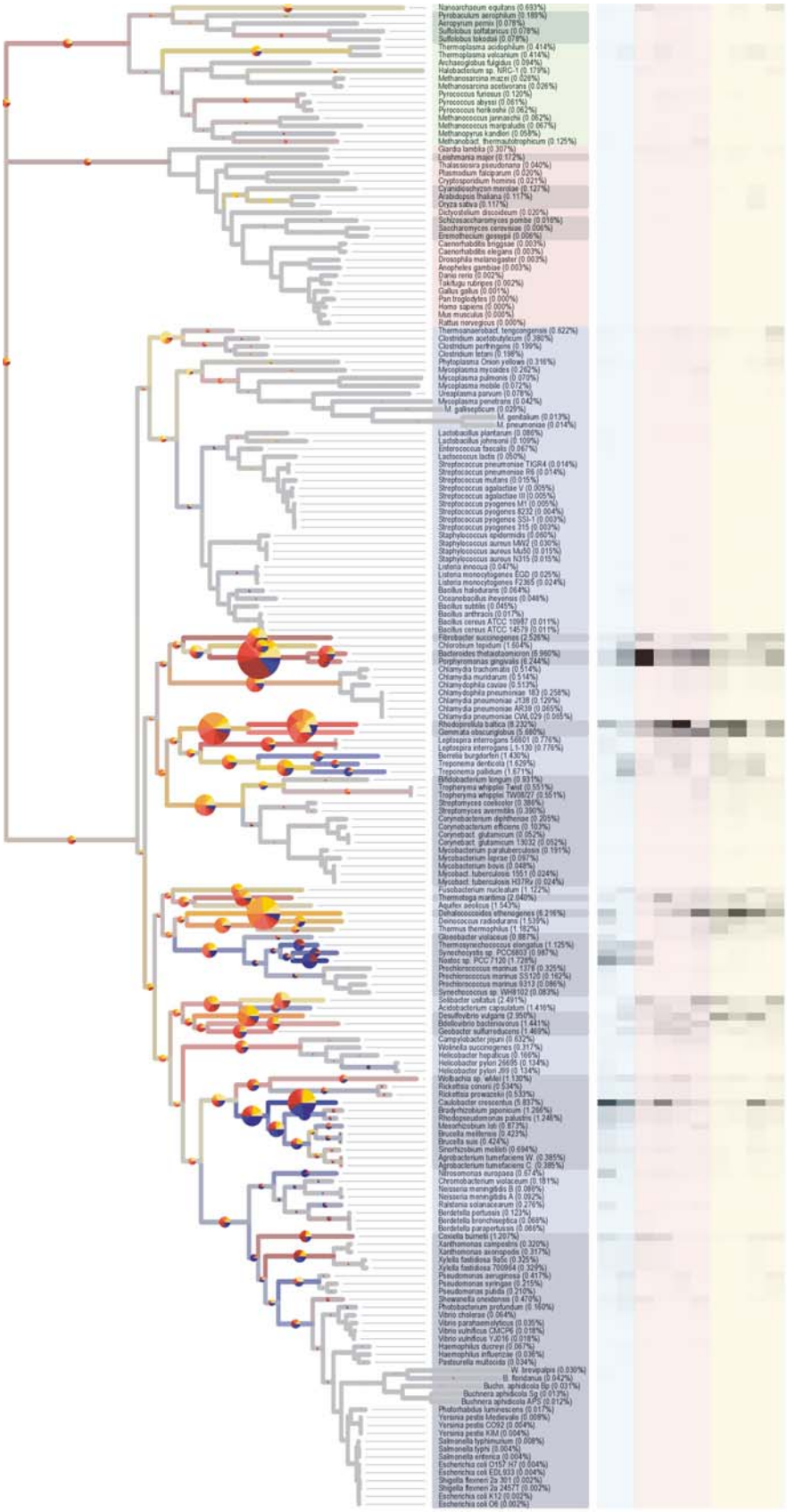
sequenced taxonomically characterized microbial isolates facilitating phylogenetic assignment of mat sequences. The phylomapping uses Maximum Likelihood mapping of phylogenetically informative genes on the tree of life to estimate the diversity of the sample (von Mering *et al*, 2007a). Each layer was mapped to a tree of 191 known genomes (Ciccarelli *et al*, 2006).

In the upper two layers, Cyanobacteria and Alphaproteobacteria were the most abundant major lineages, their abundance sharply decreasing with depth. This is consistent with the steep reduction in oxygen concentration and light intensity through the dense black-colored mat (Jorgensen and Des Marais, 1988). However, genes mapping to cyanobacteria were still readily detectable throughout the mat. Below the upper two layers, BLAST-based phylotyping was complicated by the fact that most reads had no significant similarity to existing isolate genomes consistent with the extreme phylogenetic diversity of the mat (Ley *et al*, 2006) and severely biased and undersampled reference microbial genome dataset. Of the reads that did have similarity to reference genomes, the Alphaproteobacteria, Gammaproteobacteria, Deltaproteobacteria and Bacteroidetes were the most represented lineages throughout the mat. Other groups identified as predominant constituents in the mat by 16S survey, namely the Chloroflexi and Planctomycetes (Ley *et al*, 2006), were not well resolved by BLAST-based analysis due to a lack of reference genomes for these phyla. Below the two uppermost layers, phylomapping identified Alphaproteobacteria, Bacteroidetes, Chloroflexi and Planctomycetes as the most abundant lineages in the mat. In addition both analyses identified traces of other bacterial phyla as well as some Archaea and eukaryotes. Overall, our findings are in agreement with previous phylogenetic analyses of this mat (Ley *et al*, 2006; Spear *et al*, 2003), which found similar gross phylogenetic distribution patterns.


A
fraction of phylogenetic assignments per layer (%)
0
5
10
15
20
25
30
35
40
0
10
20
30
40
Depth (mm)
Unassigned
Alphaproteobacteria
Cyanobacteria
Deltaproteobacteria
Gammaproteobacteria
Bacteroidetes


B
fraction of phylogenetic assignments per layer (%)
0
5
10
15
20
25
30
35
40
45
50
0
10
20
30
40
Depth (mm)
Alphaproteobacteria
Bacteriodetes
Chloroflexi
Cyanobacteria
Planctomycetes

C



100

Figure S1. Phylogenetic characterization of the mat by layer. (A) Count of best hits by phylogenetic group using a 30% BLAST identity threshold against the IMG database. Only the most abundant phyla and proteobacterial classes are shown. Genes with hits below the threshold, or without any hits are unassigned. (B) Phylogenetic mapping of the 10 mat layers on the Ciccarelli et al. tree. (Ciccarelli *et al*, 2006), only dominant groups are shown. Sequences were mapped using the method of (von Mering *et al*, 2007a). In short, for each sample, predicted proteins that belonged to a set of phylogenetic marker orthologous groups were aligned to a set of hand-curated alignments. From these alignments, metagenomic proteins were mapped to the reference tree using maximum likelihood (see (von Mering *et al*, 2007a) for details). For the broader groupings shown here, mappings for that subgroup were summed and the fraction of each group to the sample total was plotted. (C) Tree-based view of (B) enabling higher phylogenetic resolution of results. On the right, intensity map of mappings, with each column being a layer (ordered from left (top) to right (deepest). For each row (tree leaf), intensity was summed over the branches leading to the leaves (with equal partition on bifurcations) and was normalized over total (i.e. the darkest black is the largest number of mappings of *all* layers - not only that layer). Colored bands (blue, red, yellow) indicate three chemical zones (oxic, lowH2S, highH2S). On the left, mappings as placed on the tree. Colored pie charts show precise mapping on branches (intensity map shows values summed over branches) - shades of blue, red, yellow indicate layers. Colors of branches indicate dominant layer for mappings on that branch. Normalized raw data for Fig. S1 are available as a linked file to the supporting information.

Average genome size

Using a method based on the fraction of single-copy marker genes (Raes *et al*, 2007), the average effective genome size (EGS) of organisms living in the mat was calculated. Two measurements were taken: 1) general EGS, which is measured from single-copy marker genes/nucleotide density and represents the average genome size over all organisms (incl. eukaryotes) accounting for multiple copies of plasmids and inserted sequences, as well as for associated phages and viruses, and 2) bacteria-specific EGS, which is measured from

the single-copy bacterial marker genes/all bacterial genes ratio and measures EGS specifically for the bacteria in the sample (see Figure S2).

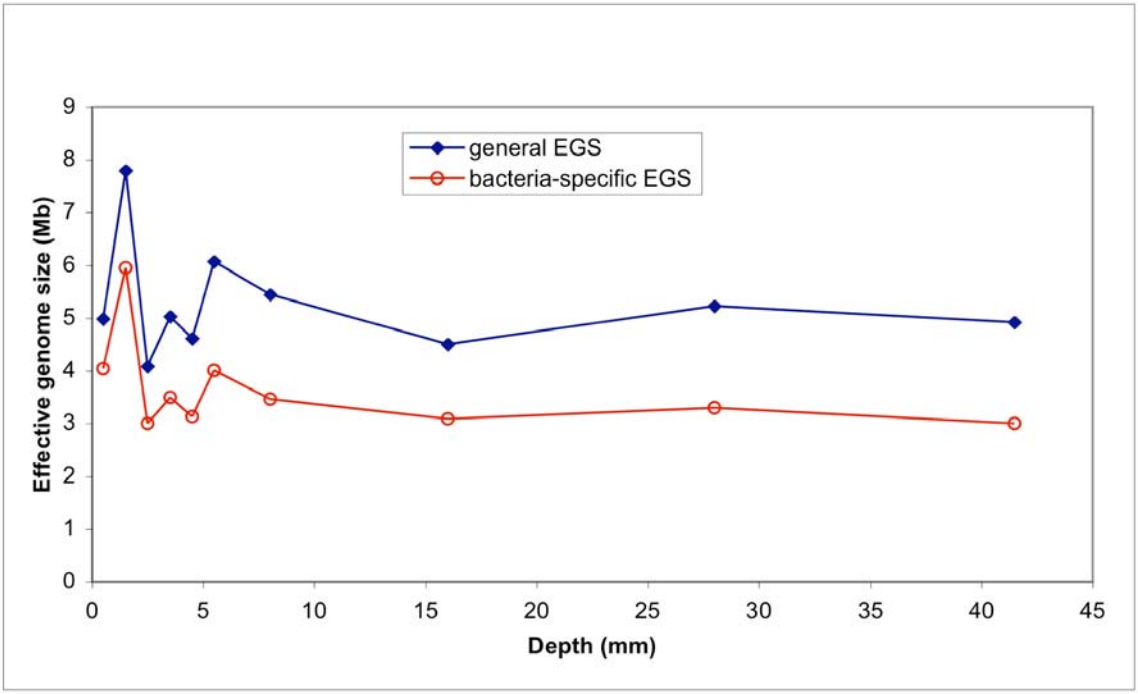


Figure S2: Effective genome size (EGS) general (blue) and bacteria-specific (red) per layer (x-axis) in Mb.

All layers except layer 2 are statistically indistinguishable and converge on an average bacteria-specific EGS of about 3-3.5Mb, which is similar to that found for other mats (Acid Mine Drainage EGS = 3.2; Whalefalls = 3.5), but much smaller than soil (Minnesota sample = 4.7) and bigger than sea (Sargasso sample = 1.6).

Interestingly, the EGS in layer 2 is significantly larger than in the other layers, both for the general and bacteria-specific measure. One could speculate that genomes in this layer are larger because of an expansion of the gene repertoire: this layer is probably the chemically most complex – strong O2 gradient, pH gradient and on the border of the oxic and H2S zone – which would necessitate an expansion to cope with a broad scale of environmental pressures.

Gene family coverage in mat

To estimate how much of the functional space of each layer was sampled using the shotgun sequencing, reads were assigned to STRING (von Mering *et al*, 2007a) orthologous groups using a 60 bit cut-off as described in (Tringe *et al*, 2005). For each sample, reads were randomly selected without replacement and total number of assigned COGs per basepair sampled was determined and plotted (Fig S3). Coverage appears to be rather low, with none of the layer curves reaching saturation. Layers 1 and 2 seem to follow a different behavior, in line with the higher environmental variability and/or complexity of these layers.

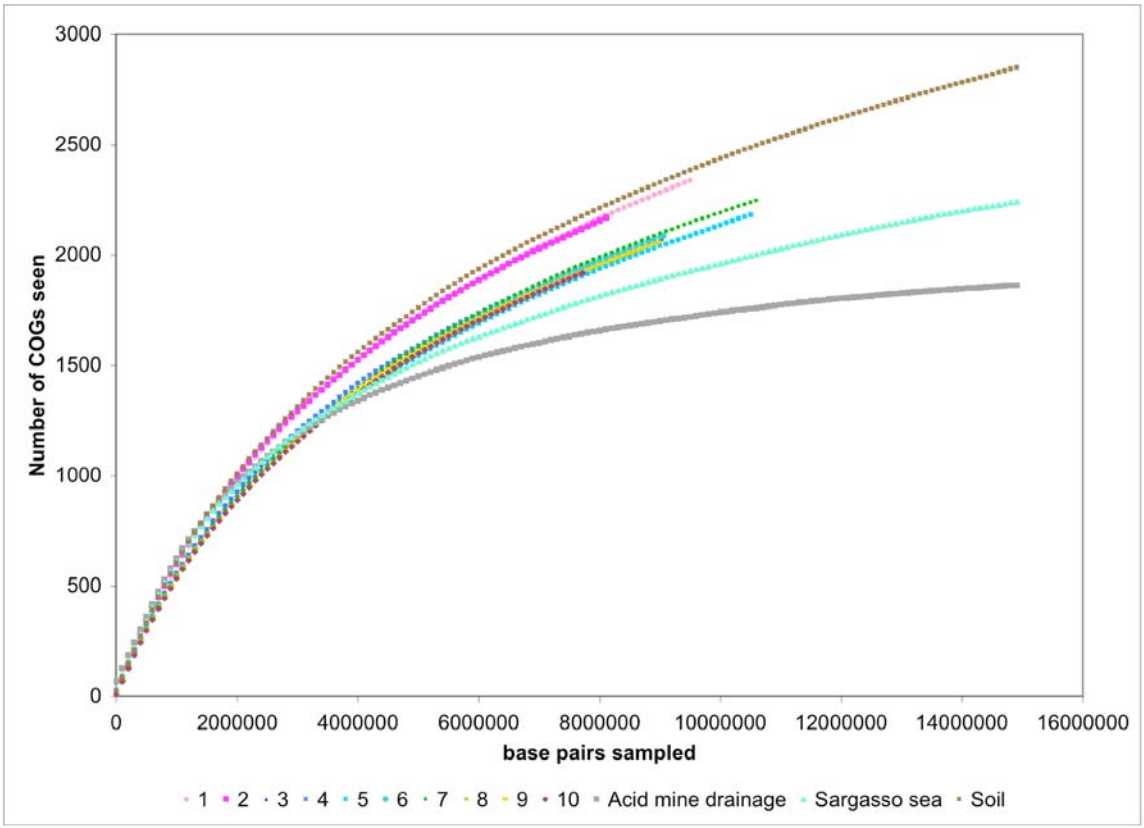


Figure S3. Collector's curves for the 10 mat layers (numbered 1 to 10) and reference datasets: acid mine drainage (Tyson *et al*, 2004), Sargasso sea (Venter *et al*, 2004) and soil (Tringe *et al*, 2005). The figure shows a Lowess fit on 10 repetitions of each sampling run (smoothing factor 0.1 using the lowess function in the R package (www.r-project.org)).

Gene-based functional gradients

The identification of COG families and pfam profiles was based on the IMG/M (Markowitz *et al*, 2006) database, using a cut-off of 20% identity and e-value 0.01. Gene family gradients were initially identified using the find functions and abundance profile tools in IMG/M. Groups of protein families were created by manually including all the possible protein families that have the function of interest while excluding protein families that are known to include proteins with alternative functions or participate in other pathways. For example, if a protein family was only known to participate in methylation associated with chemotaxis, it was included in the chemotaxis category, while a protein family that can do methylation either associated or not associated with chemotaxis was excluded.

The significance of the observed gradients was estimated by 1,000 bootstrap samplings of predicted proteins as follows. In each bootstrap iteration of proteome sized N, N predicted proteins were selected randomly from the dataset, with multiple samplings of the same protein allowed. In this bootstrapped dataset the number of occurrences of the protein family (or group of protein families) of interest was recorded. After 1,000 bootstrap iterations, an array of 1,000 observations was created. A standard deviation was computed from this array using Math::NumberCruncher perl module to provide confidence estimates for datapoints in Figure 1. Values on the graph with non-overlapping standard deviations were considered as significantly different. See main text for results and discussion.

Gradients were independently verified using the STRING database (version 7.0; (von Mering *et al*, 2007b)) as follows. Proteins with a BLAST bit score>60 were mapped onto COGs, operons and KEGG maps using the same procedure as in (Tringe *et al*, 2005). To detect gradients, the following combinations of pooled layers were tested for significant (e-val<0.05) overrepresentation of COGs, operons and KEGG maps (Tringe *et al*, 2005) (Tringe et al.) in either pool: Oxic (layers 1,2) vs. low H2S (layers 3,4,5,6) vs. high H2S (layers 7,8,9,10); Oxic vs. low+highH2S , top half of layers (1-5) vs. bottom half (6-10). Significant classes (below) were then checked manually to eliminate artifacts (e.g. incomplete KEGG maps) and/or provide more detailed explanations of observed trends (see main paper).

Component COGs or pfam domains of functional groups presented in Fig. 1.

Ferredoxins and associated proteins

COG0348, COG0633 COG0674, COG1013, COG1014, COG1018, COG1139, COG1141, COG1144, COG1146, COG1148, COG1149, COG2146, COG2414, COG2440, COG3411, COG4231, COG4739, COG4802

Sugar degradation

COG0149, COG0191, COG0205, COG3588, COG1312, COG1904, COG2160, COG2407, COG4806, COG2115, COG0524, COG1070, COG1082, COG0036, COG0800, COG2721, COG3717, COG3734, COG3954.

Chaperones

COG0071, COG0326, COG0443, COG0484, COG0501, COG0533, COG0542, COG0576, COG0606, COG1214, COG1281.

Photosynthesis

pfam03437, pfam05447, pfam07143, pfam02276, pfam06206, pfam06485, pfam07082, pfam05969, pfam05996, pfam03130, pfam01716, pfam00124, pfam03967, pfam02605, pfam00737, pfam02532, pfam01788, pfam02533, pfam02419, pfam02468, pfam01789, pfam01405, pfam03912, pfam06596, pfam06298, pfam00796, pfam01701, pfam00421, pfam05398, pfam02392.

Chemotaxis

COG0840, COG0643, COG2201, COG1352, COG0835.

Flagella

COG1157, COG1256, COG1261, COG1291, COG1298, COG1317, COG1334, COG1338, COG1344, COG1345, COG1360, COG1377, COG1419, COG1516, COG1536, COG1558, COG1580, COG1677, COG1681, COG1684, COG1705, COG1706, COG1749, COG1766, COG1815, COG1843, COG1868, COG1886, COG1987, COG2063, COG2874, COG2882.

Isoelectric point and amino acid composition

Isoelectric point was calculated for each protein in the mat and reference genome or microbiome dataset (obtained from the IMG database version 2) using pICalculator from the Bio::Tools module of bioperl. The average isoelectric point was calculated from these data for each mat layer and reference dataset (Figure 1A). The mat layers had very close average isoelectric points ranging from 6.5 to 6.7, which is lower than the microbial average (7.3; Table S3). This is not an artifact of the metagenomic data, as other metagenomic datasets, some processed by an identical pipeline at JGI, are much closer to the microbial average.

Table S3. Average isoelectric points of selected metagenomic projects.

Sample		Average isoelectric point
Mat layers	1	6.50
	2	6.60
	3	6.69
	4	6.55
	5	6.56
	6	6.58
	7	6.57
	8	6.51
	9	6.52
	10	6.55
Whalefall Sample #1		6.95
Whalefall Sample #2		6.98
Whalefall Sample #3		7.12
Sludge US		7.62
Sludge Australian		7.35
Olavius spp. symbionts		7.65
Acid Mine Drainage		7.68
Soil		7.74
All microbial genomes		7.3

We also plotted the isoelectric point (Fig. 3A) and GC content (Fig. 3B) profiles for each mat layer. Bacterial and archaeal isoelectric point profiles were described to be bimodal (Schwartz *et al*, 2001) with cytoplasmic proteins comprising the left (more acidic) peak and membrane proteins comprising the right (more basic) peak. Organisms with a salt-in strategy such as *Salinibacter* have acid shifted peaks with a higher ratio of low isoelectric

point (left) to high isoelectric point (right) peaks. The isoelectric point profiles for the mat are consistent with a high contribution of salt-in halophiles (Fig. 3A). Moreover, the profiles of the mat layers were almost identical despite variable GC content (Fig. 3). To identify why the mat protein isoelectric points were acid-shifted, we wrote a perl script to calculate the content of each amino acid in the mat proteins as compared to amino acid usage of reference microbes and microbiomes. The most striking difference was in the abundance of aspartate which is remarkably homogenous between mat layers and higher than most other bacteria, archaea and metagenomes (Fig. 2B). The usage of aspartate is not influenced by GC content since it can be encoded by both GC-rich and GC-poor codons consistent with the variable GC content between mat layers (Fig. 3B). The use of higher aspartate content in proteins to enhance their acidity has only been reported for salt-in halophilic archaea which do not feature prominently in the mat community.

We also investigated the dataset for genes associated with the salt-out strategy, i.e. genes encoding compatible solutes (Wood, 2007). Glycine betaine transporters (COG0834, COG0765, COG1126 COG1125, COG1174, COG1292, COG1732, COG2113, COG4175, COG4176) are present in the mat but occur in low numbers. Indeed, the same gene families and are present in higher abundance in other habitats, including soil and gutless worm symbionts (data not shown). Similarly, we found only traces of the ectoine synthesis pathway (pfam06339). While it is highly likely that the salt-out strategy is employed by mat halophiles, we have insufficient molecular data to discern gradients within the mat. There are several possible explanations for the low incidence of salt-out genetic traits; 1) coverage is too low to identify abundance gradients of these low copy number protein families, 2) there is no increased abundance of gene families, but increased expression of relevant proteins, 3) novel compatible solutes are used or 4) the salt-out strategy is simply not widely used by the Guerrero Negro mat community.

References

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, NY* **311:** 1283-1287.

http://www.softberry.com/ SoftBerry - fgenesb.

Jorgensen BB, Des Marais DJ (1988) Optical properties of benthic photosynthetic communities: fiber-optic studies of cyanobacterial mats. *Limnol Oceanogr* **33:** 99-113.

Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, Maresca JA, Bryant DA, Sogin ML, Pace NR (2006) Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol* **72:** 3685-3695.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavrommatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P, Kyrpides NC (2006) An experimental metagenome data management and analysis system. *Bioinformatics (Oxford, England)* **22:** e359-367.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* **8:** R10.

Schwartz R, Ting CS, King J (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome research* **11:** 703-709.

Spear JR, Ley RE, Berger AB, Pace NR (2003) Complexity in natural microbial ecosystems: the Guerrero Negro experience. *The Biological bulletin* **204:** 168-173.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. *Science (New York, NY* **308:** 554-557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, NY* **304:** 66-74.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P (2007a) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science (New York, NY* **315:** 1126-1130.

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007b) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic acids research* **35:** D358-362.

Wood JM (2007) Bacterial osmosensing transporters. *Methods in enzymology* **428:** 77-107.